

El Metodo Kimball – Integración de Datos

La metodología de Kimball, llamada Modelo Dimensional (Dimensional Modeling), se basa en lo que se denomina Ciclo de Vida Dimensional del Negocio (Business Dimensional Lifecycle). Esta metodología es considerada una de las técnicas favoritas a la hora de construir sistemas Data Warehouse y Business Intelligence.



Figura 1: Ralph Kimball

Un Data Warehouse es una base de datos corporativa, centralizada que contiene datos y metadatos denominados objetos, cuyo proposito es el desarrollo de procesos analiticos y de consultas variadas basados en la implementacion de modelos o estructuras multidimensionales (multitabulares) tambien denominadas cubos o datamarts y cuyos resultados sirven para el apoyo a la toma de decisiones.

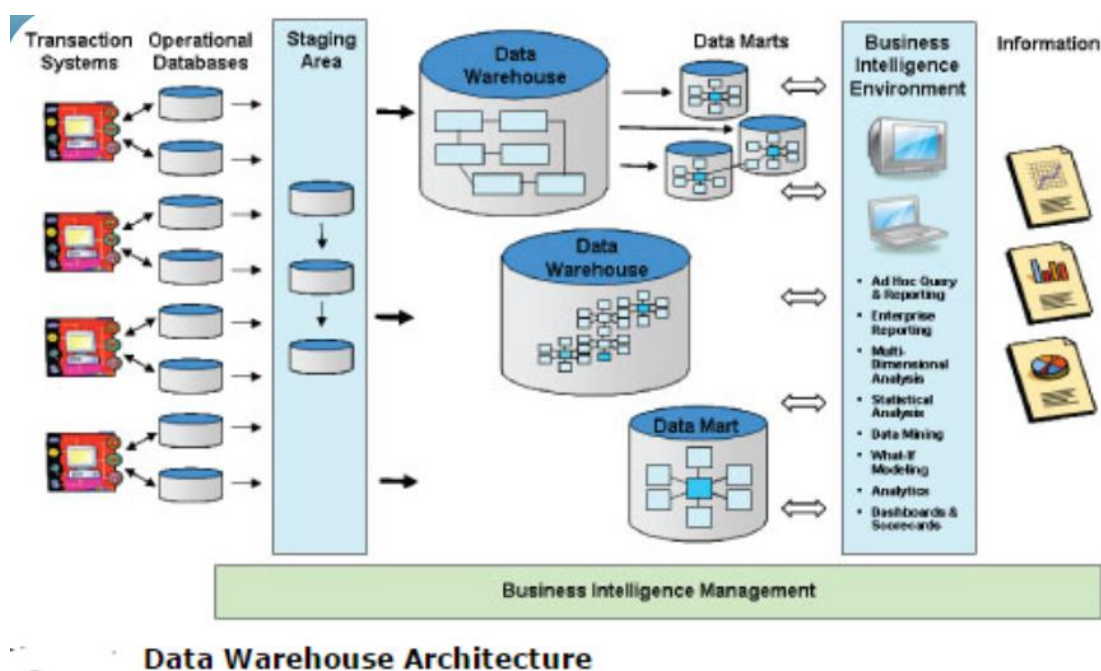


Figura 2: DataWarehouse

Kimball recomienda algunos principios a tener en cuenta para la creación de los datawarehouses empresariales. Estos principios son:

1. Seguir una metodología probada (se recomienda la metodología del ciclo de vida Kimball).
2. Comprender con mucha claridad los requerimientos del negocio para poder traducirlos en un modelo de datos, priorizando los esfuerzos y dando valor agregado a la organización.
3. Diseñar las áreas de datos del DW de modo de hacerlo flexible, reusable y con alto nivel de performance.
4. Implementar en forma rapida y progresiva incrementos basados en procesos de negocios que conforman la matriz de datos empresariales tambien denominada como Matriz BUS de DW.
5. Diseñar una arquitectura DW que responda a los procesos del negocio, a los volúmenes de datos y a la infraestructura de TI.
6. Construir la solución ETL con componentes estandars para poder manejar los modelos de diseño estandares que se utilizan en las plataformas de analítica de datos externos.
7. Entregar una solución completa que incluya reportes, programas de consulta, portales, documentacion, entrenamiento y soporte.

*

El Ciclo de Vida de Kimball es una metodología detallada para el diseño, desarrollo e implementación de sistemas de BI y DW. La figura 1 muestra las etapas principales de la metodología.

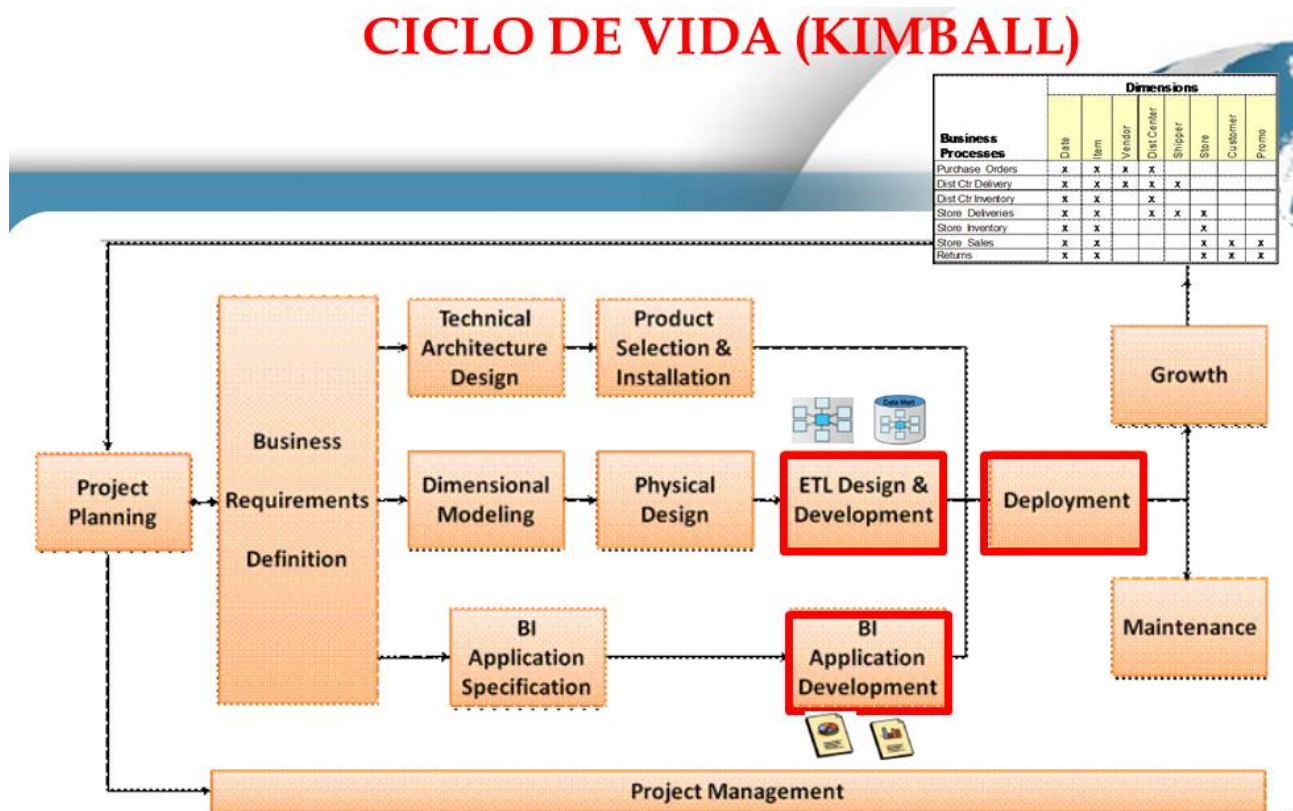


Figura 3: Ciclo de Vida Kimball

El ciclo de vida es un enfoque iterativo, en la que en cada pasada se generan un conjunto coherente de datos y estructuras de datos y un conjunto de informes y aplicaciones analíticas asociadas. Cada ciclo de iteración puede

ser terminado aproximadamente en un período de 6 a 9 meses, dependiendo de la complejidad de los procesos de análisis y datos. La implementación del sistema DW/BI puede tomar múltiples iteraciones, cada una de las cuales carga nuevos contenidos de datos (objetos de datos, estructuras multidimensionales, aplicaciones de BI) que están relacionadas con la matriz de datos de la empresa denominada bus matrix, que representa el modelo de análisis del Sistema de BI/DW.

*

La metodología Kimball comienza con la comprensión de los requerimientos del negocio y como se puede dar valor a la organización a través de la información. La actividad inicial debería comenzar por entrevistas para determinar las prioridades de información de la organización, teniendo como resultado una lista ordenada de procesos de negocios que producen datos, el ámbito del proyecto, la granularidad de la información, las dimensiones de datos (objetos de datos) y los hechos que serán analizados, junto con los procesos analíticos de alto valor para la empresa que pueden desarrollarse con dichos datos. De este análisis podemos ya construir el modelo conceptual preliminar del DW de la empresa y obtener la Matriz BUS de la empresa base para el diseño del sistema DW/BI.

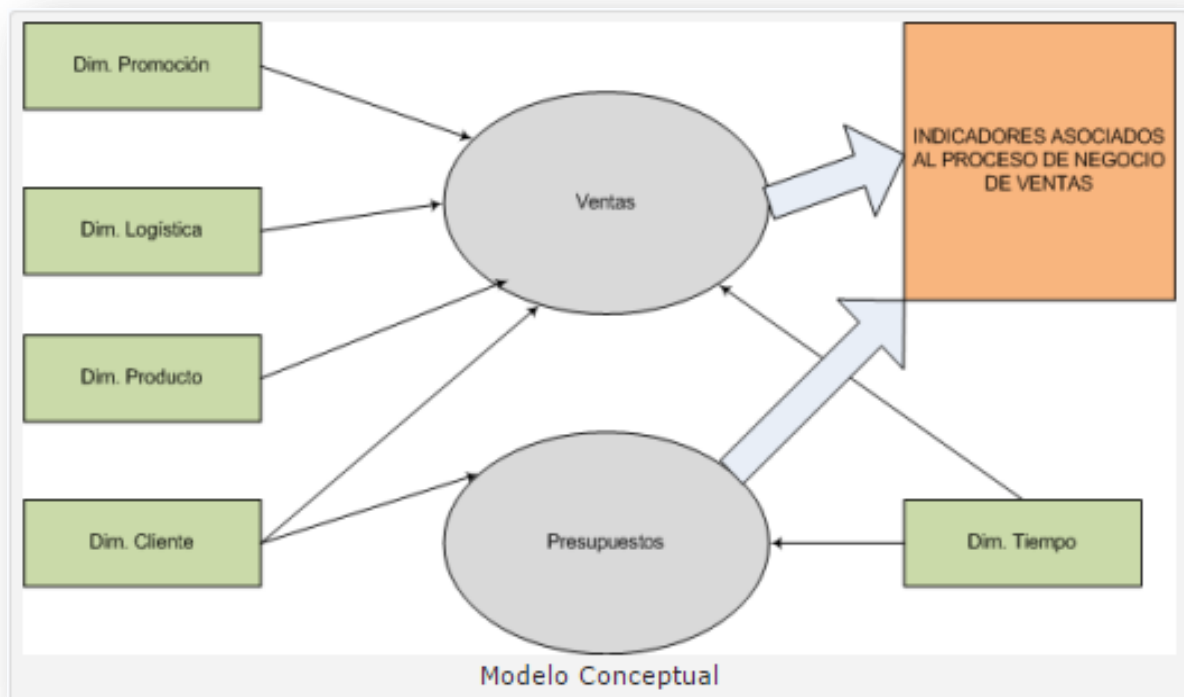


Figura 4: Modelo Conceptual

*

Business Processes	Dimensions							
	Date	Item	Vendor	Dist Center	Shipper	Store	Customer	Promo
Purchase Orders	X	X	X	X				
Dist Ctr Delivery	X	X	X	X	X			
Dist Ctr Inventory	X	X		X				
Store Deliveries	X	X		X	X	X		
Store Inventory	X	X				X		
Store Sales	X	X				X	X	X
Returns	X	X				X	X	X

Figura 5: Matriz BUS

El siguiente paso es tomar a partir de la matriz BUS en forma ordenada cada proceso de negocio (empezando por el de mayor prioridad o importancia) y a continuación detallar los requerimientos de negocio relacionados con cada proceso, focalizando el esfuerzo en identificar los datos y las fuentes de datos, que incluyen atributos, definiciones, reglas de negocio, standares de calidad, y el conjunto de aplicaciones analíticas que serán desarrolladas para aprovechar estos datos.

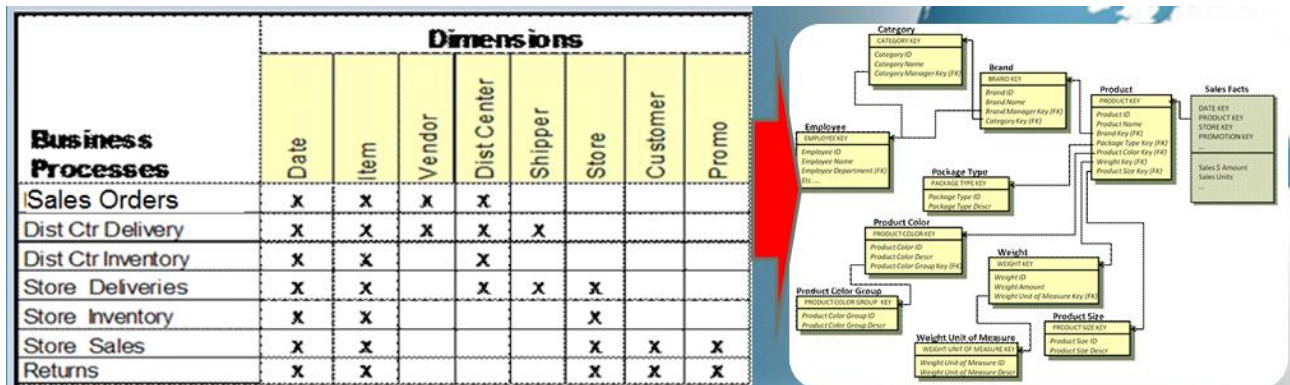


Figura 6: Desarrollo de requerimientos

Una vez terminada la etapa anterior, el ciclo de vida comienza la etapa de diseño por medio de 3 vías paralelas. La primera vía en la parte superior es la del diseño de la plataforma tecnológica, cuyo objetivo es identificar las herramientas necesarias que van a satisfacer el requerimiento del negocio.

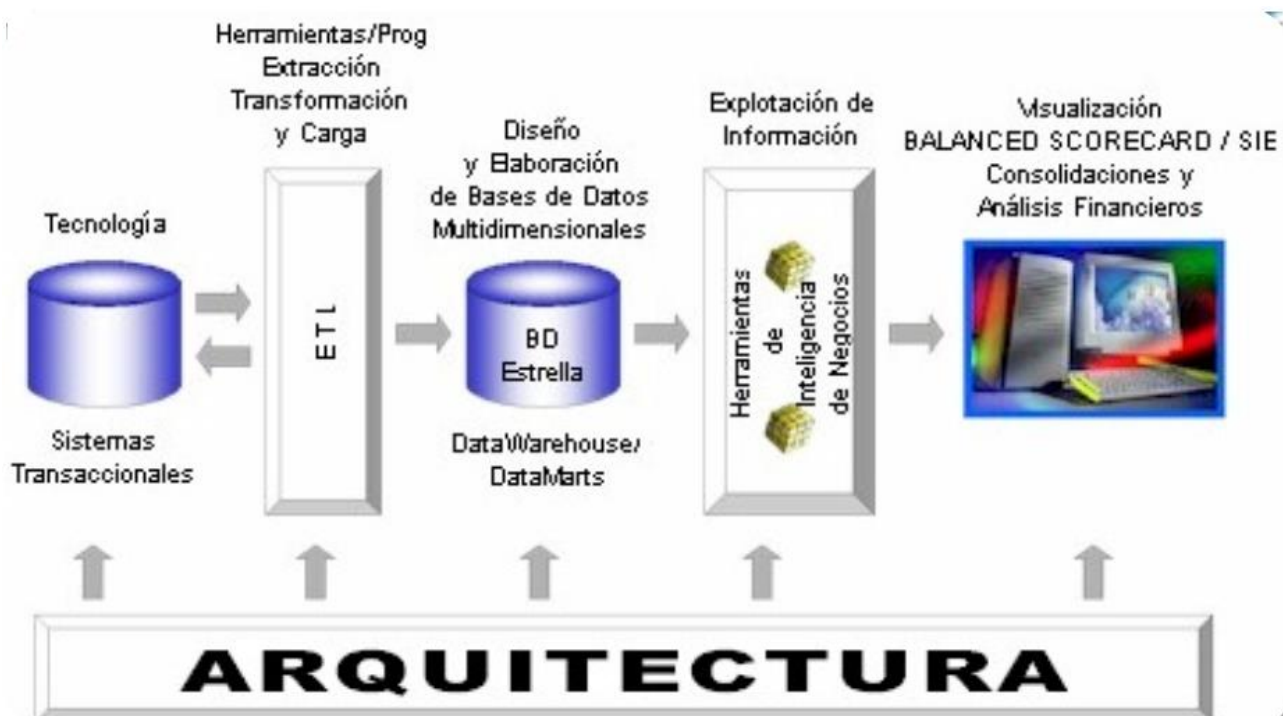


Figura 7: Definición de la Plataforma Tecnológica de Desarrollo

La vía intermedia es la del diseño de datos, que comienza con la definición del modelo de datos lógicos que satisface los requerimientos del negocio y que según Kimball define el modelo dimensional.

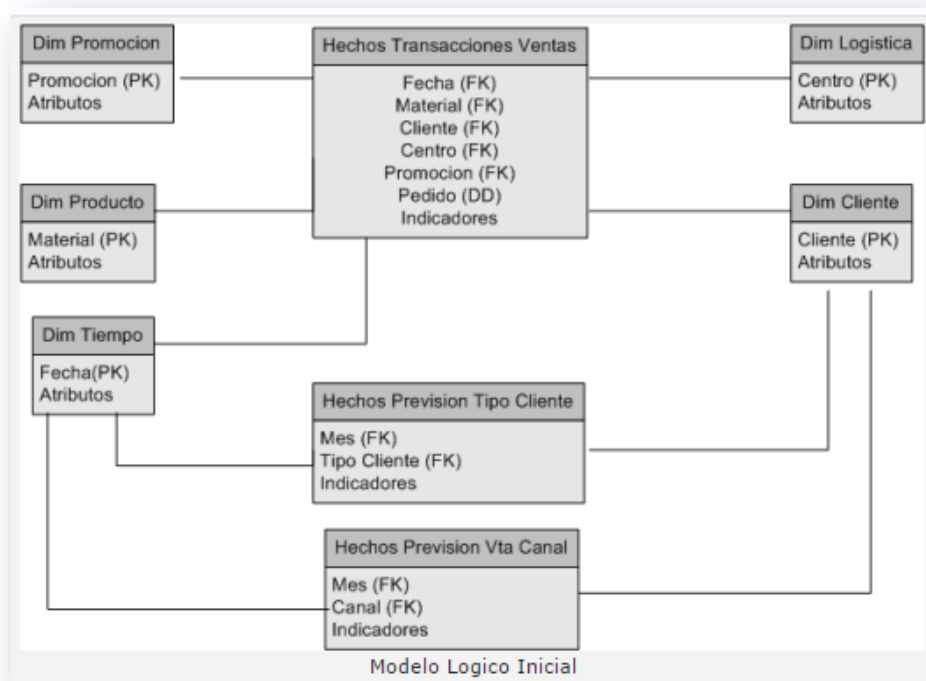


Figura 8: Modelo de Datos Lógico

Cuando el modelo lógico esta terminado, el equipo puede construir las estructuras destino en la base de datos. Las características del modelo físico dependen de la plataforma de destino a usar. El ultimo paso es crear el sistema ETL que va a cargar las bases de datos de destino. *El sistema ETL consume una gran cantidad de recursos iniciales del proyecto.

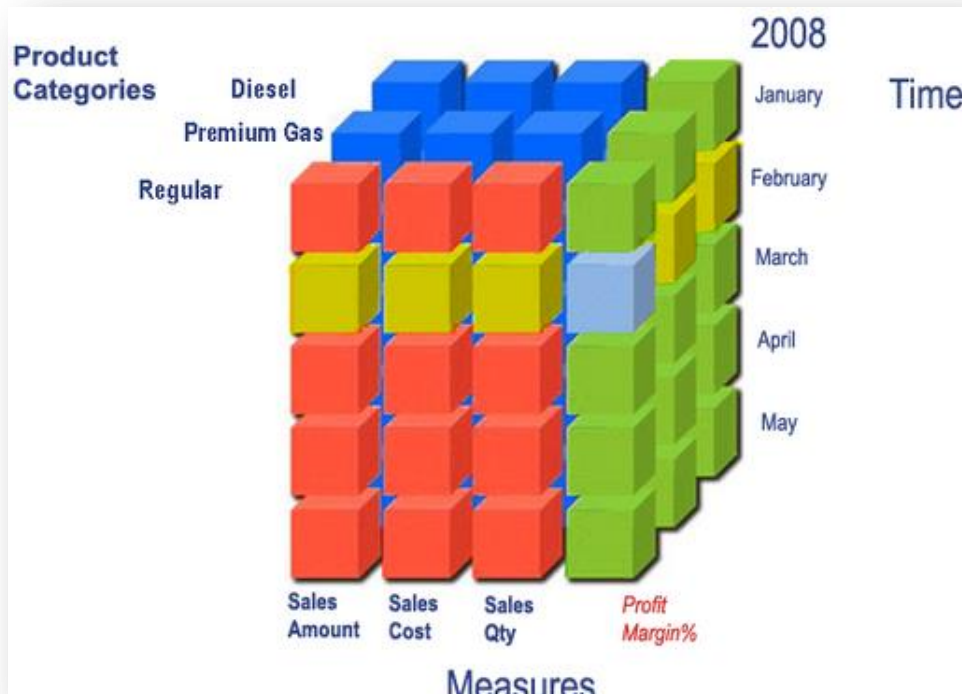


Figura 9: Modelo de Datos Físico

La tercera via crea las aplicaciones de BI, reflejados por los reportes y aplicaciones analíticas que crean valor en la empresa. Esta via se divide en 2 partes; la primera parte es la etapa de diseño donde se definen las aplicaciones analíticas y reportes de una manera detallada, en tanto la segunda parte corresponde a la implementación de estas aplicaciones analíticas y reportes. Esta segunda parte en forma general debe esperar que el proceso ETL sea implementado y los datos esten disponibles en la base de datos corporativa.



Figura 10: Implementación de aplicaciones analíticas

Cuando las tres vias se han completado, se implementa la solución de aplicaciones analíticas, reportes y consultas en forma conjunta a la comunidad de usuarios, a través de entrevistas, entrenamientos, documentacion y soporte.

La siguiente iteración (para el siguiente proceso de negocios) comienza generalmente durante la implementación del proceso previo, en el momento que los analistas y diseñadores han recolectado todos los requerimientos para el siguiente proceso de negocios de mas alta prioridad, a continuación se debe crear el modelo de dimensiones asociado iniciando el proceso nuevamente y repitiéndose el ciclo. El metodo incremental del ciclo de vida es un elemento fundamental que genera valor a la empresa en poco tiempo, construyendo paralelamente una plataforma de información empresarial de largo plazo.

La Matriz BUS de DataWarehouse

Es una red de datos ligada al diseño del DataWarehouse Empresarial, donde los encabezados en la parte lateral izquierda representan los procesos de negocios principales de la organización. Estos procesos pueden estar referidos a los procesos que conforman la cadena de valor y que son los que elaboran los productos y servicios que se entregan al mercado.

Las cabeceras de las columnas en la matriz representan los objetos primarios que conforman los procesos de negocios, como por ejemplo: cliente final, proveedor, promocion, producto, almacen, empleado y fecha. Estos objetos se denominan dimensiones y deben estar integrados para ser utilizados con todos los otros procesos importantes de negocios.

Enterprise Bus Matrix											
Adventure Works Data Warehouse Bus Matrix	Business Priority	<-- Conformed Dimensions -->									
		Date (Order, Start, Ship)	Product	Promotion	End Customer	Employee	Reseller	Page	Internet Registered User	Part	Vendor
Business Process											
Orders Forecasting	2	x	x	x		x	x				
Reseller Orders	1	x	x	x		x	x				
Internet Orders	1	x	x	x	x			x	x		
Purchasing		x	x		x	x				x	x
Parts Inventory		x	x	x						x	x
Manufacturing	6	x	x							x	
Finished Goods Inv.		x	x	x							
Shipping	3	x	x	x	x	x	x				x
Returns	5	x	x		x	x	x				x
Customer Calls	4	x	x	x	x	x	x			x	
Web Support	4	x	x		x	x	x	x	x		x

Key Concepts:

- The high level DW/BI data architecture
- Rows = Business Processes
- Columns = Conformed Dimensions
- DW/BI system implemented row by row based on business priority

Figura 11: Matriz BUS

Esta integración denominada conformación de datos, estandariza nombres, descripciones, mapeo, jerarquias y reglas de negocios que se aplicaran al sistema DW/BI. Este proceso lo realiza el sistema de administración de datos de la organización o MDM. Una vez terminada la estandarización, las dimensiones pueden ser reutilizadas en todos los procesos de negocios asociados y aun mas importante que esto las dimensiones formadas son las estructuras requeridas para la integración y donde los resultados de 2 o mas procesos pueden ser combinados para obtener una salida en la plataforma de BI.

Cada fila de la Matriz es una area de datos asociada a un proceso de negocios y corresponde a una unidad de trabajo que debe ser implementada en el sistema ETL. Cada area de datos asociado a un proceso de negocios requiere un modulo dedicado ETL para preparar las tablas de hechos, juntamente con las dimensiones asociadas y unir las en un unico modelo dimensional.

Diseño del Modelo de Datos

En general casi todas las estructuras multidimensionales se basan en el modelo estrella, donde los valores numericos asociados a los procesos de negocios estan concentrados en una tabla de valores (hechos) en la parte central del modelo, y la información contextual referida a dichos valores numéricos viene representada por un conjunto de tablas de dimensiones denormalizadas que estan alrededor de la tabla central de hechos.

Los campos claves de union entre las tablas de hechos y las tablas de dimensiones deberian ser valores enteros anonimizados denominados claves surrogadas.

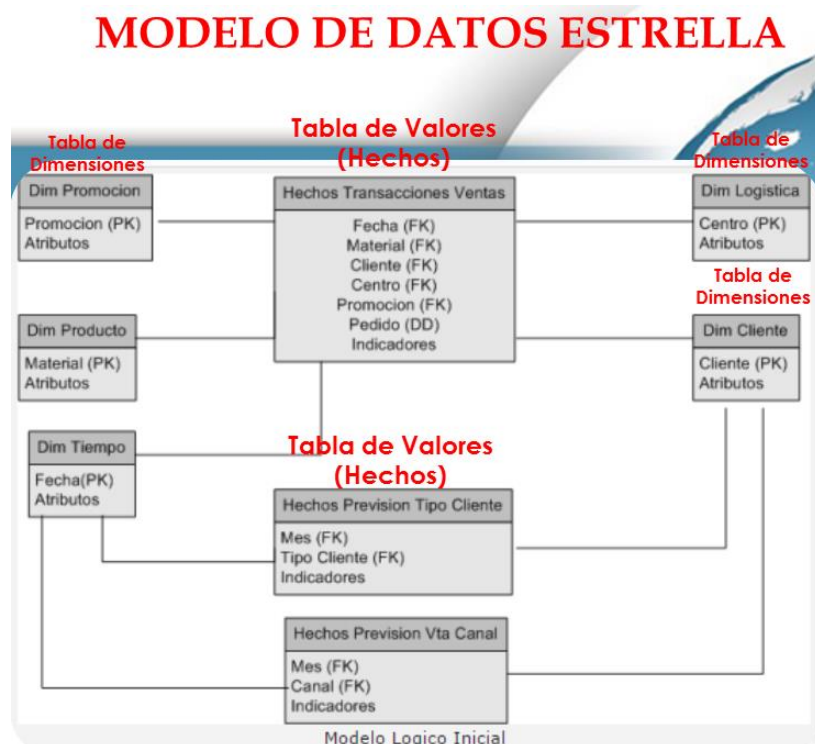


Figura 12: Modelo Estrella

Usabilidad

Todos los componentes de datos contenidos en el Datawarehouse estan adaptados para formar parte de las estructuras multidimensionales que son de facil acceso a todos, que según estudios son modelos de datos de facil entendimiento y uso en comparación a otros modelos normalizados.

Flexibilidad

Existe una escuela de pensamiento que sostiene que la tercera forma normal de los datos viene detallada en los datos atomizados del datawarehouse, que proporcionan una gran flexibilidad al modelo. Pero si bien esto puede ser cierto desde una perspectiva de sistemas transaccionales, se debe tener en cuenta que se esta construyendo una base de datos analítica.

La mayor parte de las transacciones de sistemas estan basadas en modelos de datos en la tercera forma normal con transacciones a nivel de detalle capturadas en tablas de hechos normalizadas. Esta escuela sostiene que estos modelos de datos son la base de los sistemas data warehouse. Por consiguiente estas estructuras requieren transformaciones adicionales para que puedan ser presentadas y consumidas por el usuario final lo que frecuentemente implica crear estructuras fisicas como de datos departamentales como los Data Marts.

El modelo normalizado y el modelo multidimensional atomizado adecuadamente diseñado son relacionalmente equivalentes y pueden responder el mismo grupo de consultas analíticas.

*

La flexibilidad del modelo depende del nivel detalle al que el modelo multimensional puede descender. Otro error de conceptualización es creer que el modelo dimensional esta formado de valores agregados o resumidos unicamente, cuando el objetivo del diseño ha sido capturar datos al nivel mas bajo de detalle disponible, denominado nivel atomico.

La datos atomizados del DataWarehouse permiten al usuario desplazarse de arriba hacia abajo en todos los niveles de datos permitidos. Cualquier agregación anterior a la carga de datos en el DW significa que el detalle de dichos datos no estaran disponibles, reduciendo asi la flexibilidad del modelo de datos.

Performance y Mantenimiento

El modelo multidimensional mantiene la tabla central de hechos en una manera normalizada (manteniendo las tablas de dimensiones fuera de la tabla de hechos en forma normalizada) para un mejor performance.

Se debe notar que las tablas de dimensiones planas contienen la misma información que las tablas de dimensiones normalizadas y que no implementan tablas adicionales ni claves extras para su proceso de normalización. El modelo dimensional reduce el numero de tablas y uniones que se requieren en procesos analíticos, y mejora la performance de la mayoría de sistemas de base de datos.

En conclusión en un implementación ideal, se debería tener los datos en su mas bajo nivel de detalle (atomizados) dentro del DW para incrementar la flexibilidad, asi tambien las evidencias demuestran que el modelo multidimensional es el mas apropiado para la reusabilidad de los usuarios, y el modelo de datos fisicos tambien deberia ser dimensional por simplicidad y performance. La experiencia respalda lo dicho.

Dimensiones y Valores

Como se ha comentado en la Matriz Bus, las dimensiones son los objetos centrales de los procesos de la organización, y generalmente cada una genera una tabla de dimensiones. Construir una dimension en el sistema ETL implica unir varias tablas normalizadas de descripciones y jerarquias que cargan los atributos de una tabla simple. Figure 4 shows an example of typical product-related attributes in a normalized model.

La tabla base es la de productos y esta conectada a la tabla de valores Sales por medio de un campo clave de Producto. A partir de estas tablas es posible crear calculos analíticos como SUM([Sales \$ Amount]) by CategoryName, or by ProductColorGroupDescr, o cualquier otro atributo en cualquiera de las tablas normalizadas que describen el producto. Es posible pero no sencillo.

En el modelo dimensional de la tabla de Producto podriamos unir las tablas asociadas al producto (fig 4) una vez durante el proceso ETL, para producir una tabla dimensional Producto.(Fig 5)

Desde el punto de vista analítico ambos modelos son equivalentes.

La usabilidad mejora para los desarrolladores de aplicaciones de BI y usuarios ad-hoc en el modelo dimensional. En este ejemplo las 10 tablas de atributos del producto se han concentrado en 1 sola. Esta reducción de 10 a 1 en el numero de tablas enfrenta a los usuarios a grandes diferencias en usabilidad y performance. Cuando se repite este proceso en una mayor cantidad de dimensiones los beneficios son enormes.